Measuring the ability of transitive reasoning, using product and strategy information

Bouwmeester, S.; Sijtsma, K.

Published in:
Psychometrika

Publication date:
2004

Document Version
Publisher's PDF, also known as Version of record

PSYCHOMETRIKA—VOL. 69, NO. 1, 123–146
MARCH 2004

MEASURING THE ABILITY OF TRANSITIVE REASONING, USING PRODUCT AND STRATEGY INFORMATION

SAMANTHA BOUWMEESTER AND KLAAS SIJTSMA

TILBURG UNIVERSITY

Cognitive theories disagree about the processes and the number of abilities involved in transitive reasoning. This led to controversies about the influence of task characteristics on individuals' performance and the development of transitive reasoning. In this study, a computer test was constructed containing 16 transitive reasoning tasks having different characteristics with respect to presentation form, task format, and task content. Both product and strategy information were analyzed to measure the performance of 6- to 13-year-old children. Three methods (MSP, DETECT, and Improved DIMTEST) were used to determine the number of abilities involved and to test the assumptions imposed on the data by item response models. Nonparametric IRT models were used to construct a scale for transitive reasoning. Multiple regression was used to determine the influence of task characteristics on the difficulty level of the tasks. It was concluded that: (1) the qualitatively distinct abilities predicted by Piaget's theory could not be distinguished by means of different dimensions in the data structure; (2) transitive reasoning could be described by one ability, and some task characteristics influenced the difficulty of a task; and (3) strategy information provided a stronger scale than product information.

Key words: cognitive ability, cognitive strategies, dimensionality of test data, IRT models, transitive reasoning, transitive reasoning scale.

1. Introduction

1.1. Definition of Transitive Reasoning

Suppose an experimenter shows a child two sticks, A and B, which differ in length, Y, such that $Y_A > Y_B$. Next, stick B is compared with another stick C, which differs in length such that $Y_B > Y_C$. In this example the length relations $Y_A > Y_B$ and $Y_B > Y_C$ are the premises. When the child is asked, without being given the opportunity to visually compare this pair of sticks, which is longer, stick A or stick C, (s)he may or may not be able to give the correct answer. When a child is able to infer the unknown relation ($Y_A > Y_C$) using the information of the premises ($Y_A > Y_B$ and $Y_B > Y_C$), (s)he is capable of *transitive reasoning*.

1.2. Theories of Transitive Reasoning

Three general theories on transitive reasoning can be distinguished. They are the developmental theory of Piaget, information processing theory, and fuzzy trace theory. These theories propose different definitions of the transitive reasoning ability and different operationalizations into transitive reasoning tasks. Consequently, the theories lead to contradictory conclusions about children's transitive reasoning ability.

1.2.1. Developmental Theory of Piaget

According to Piaget's theory (Piaget, Inhelder, & Szeminska, 1948), children acquire the cognitive operations to understand rules of logic at the *concrete operational stage*, at about six or

Requests for reprints should be sent to Samantha Bouwmeester, Department of Methodology and Statistics FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, Phone: 31134663270, Fax: 31134663002, Email: s.bouwmeester@uvt.nl

0033-3123/2004-1/2003-1056-A $00.75/0

seven years old. This understanding implies that an object can have different relations with other objects. For example, a stick can be longer than a second stick and shorter than a third stick. This understanding is necessary to draw transitive inferences (Piaget & Inhelder, 1941; Piaget & Szeminska, 1941). At the *pre-operational stage*, before the concrete operational stage, children think in a nominal way. This means that objects are understood in an absolute form, but not in relation to other objects. Consequently, at this stage children are incapable of drawing a transitive inference.

Piaget distinguished two kinds of reasoning. To understand a transitive inference, the formal rules of logic has to be acquired and applied to the transitive reasoning problem. This kind of reasoning was called *operational reasoning*. A child is able to reason in an operational way at the concrete operational stage. However, Piaget argued that operational reasoning is not necessary in each kind of task. When some kind of spatial cue in the task gives information about the ordering of objects (e.g., when all objects are presented simultaneously), operational reasoning is not required because the information given by the spatial cue can be used to infer the transitive relation; for example, objects become smaller from right to left. In this case, no formal rules have to be understood. Piaget called this kind of reasoning *functional reasoning*. Functional reasoning is acquired at the pre-operational stage. Piaget was in particular interested in the development of logical comprehension, and therefore used transitive reasoning tasks in which the premises were successively presented to be sure that children had to reason in an operational way. When a successive presentation of the premises is used, spatial cues about the ordering of objects are not available (although other kinds of ordering cues might be available).

1.2.2. Information Processing Theory

Although within information processing theory a broad diversity of ideas about information processing exists, differently oriented researchers on transitive reasoning do not make a distinction between functional and operational reasoning. An understanding of formal logical rules is not a necessary condition for drawing transitive inferences in any version of information processing theory. For example, in their linear ordering theory Trabasso, Riley, and Wilson (1975) and Trabasso (1977) emphasized the linear ordering in which the premise information was encoded and internally represented. Linear ordering was the only ability involved in transitive reasoning, rendering it a one-dimensional construct. Task characteristics like presentation form (*simultaneous* or *successive*), task format (e.g., $Y_A > Y_B > Y_C$ and $Y_A = Y_B = Y_C = Y_D$), and content of the task (*physical*, like length; or *verbal*, like happiness) might influence the difficulty to form an internal representation, but the same ability is assumed for all kinds of transitive reasoning tasks.

Sternberg (1980a, 1980b) and Sternberg and Weil (1980) studied the development of linear syllogistic reasoning, a special form of transitive reasoning in which the premise information is presented verbally. Sternberg (1980b) showed that a mixed model, which contains both a linguistic component and a spatial component, could explain linear syllogistic test data (for alternative models, see also Clark, 1969; DeSoto, London, & Handel, 1965; Huttenlocher, 1968; Huttenlocher & Higgens, 1971; Quinton & Fellows, 1975; and Wright, 2001). According to this mixed model, both a verbal and a linear ordering ability are involved in solving linear syllogistic reasoning tasks. Premise information is first encoded linguistically, and then ordered spatially into an ordered internal representation.

1.2.3. Fuzzy Trace Theory

According to fuzzy trace theory (Brainerd & Kingma, 1985, 1984; Reyna & Brainerd, 1990), the level of exactness of encoded information varies along a continuum. One end is defined by *fuzzy traces*, which are vague, degenerate representations that conserve only the sense

of recently encoded data in a schematic way. The other end is defined by *verbatim traces*, which are literal representations that preserve the content of recently encoded information with exactitude. These verbatim traces contain information like: there is a red object and a yellow object; the objects are vertical bars; and the red bar is longer than the yellow bar. At the other end of the continuum, the information is stored in a degraded, schematic way; for example, *objects get longer to the left* (Brainerd & Kingma, 1985; Reyna & Brainerd, 1990). The various levels of the continuum process in parallel; that is, by encoding literal information from a task, at the same time degraded fuzzy information is processed at several levels. Brainerd and Kingma (1984, 1985), and also Reyna and Brainerd (1990) showed that the fuzzy end, containing degraded information about the ordering of objects, was used to draw a transitive inference.

Fuzzy trace theory does not distinguish operational and functional reasoning (Reyna & Brainerd, 1992; see also Chapman & Lindenberger, 1992). It is assumed that task characteristics influence the level of the fuzzy trace continuum that may be used and, consequently, determine the difficulty level of a transitive reasoning task. No logical rules have to be applied and one ability, which is the ability to form and use fuzzy traces, explains an individual's performance on different kinds of tasks, rendering the construct of transitive reasoning a one-dimensional construct.

1.2.4. Comparison of Theories

Number of Abilities Involved The most important point of disagreement is *what* the ability to draw a transitive inference really *is*. Piaget distinguished operational and functional reasoning, two forms of reasoning that were qualitatively different, and acquired at different stages of cognitive development. Trabasso, Riley and Wilson's (1975) linear ordering theory assumes that forming an internal representation of the objects is one ability. Sternberg, who studied linear syllogistic reasoning, assumed a mixed model in which both a verbal and a spatial ability are involved. They are assumed to function as two separate abilities. Fuzzy trace theory also assumes one ability, that is, reasoning based on a fuzzy continuum.

From the perspective of Piaget's theory, information processing theory and fuzzy trace theory define transitive reasoning as a functional form of reasoning only applicable to a limited set of transitive reasoning tasks in which a linear ordering of the objects is given by a spatial cue. This functional reasoning does not require an understanding of transitivity, which is only acquired when children are capable of operational reasoning (Chapman & Lindenberger, 1988).

Influence of Task Characteristics on Difficulty Although not all theories make explicit predictions about the influence of task characteristics on the difficulty of a task,[1] implications with respect to difficulty can be inferred from the theories' assumptions.

- *Piaget's Theory*. Firstly, because simultaneously presented tasks can be solved by functional reasoning while successively presented tasks must be solved by operational reasoning, from Piaget's theory it can be inferred that simultaneous presentation of the premises of a task is easier than successive presentation. Secondly, because the same logical rules are needed to solve equality, inequality, or mixed equality-inequality task formats, the format of the task (e.g., $Y_A > Y_B > Y_C$, or $Y_A = Y_B = Y_C$) does not influence the difficulty of a task. Thirdly, because content of the relationship does not influence the application of logical rules, type of content does not influence the difficulty level of a task. However, Piaget first used length and then other concrete observable relationships to study transitive reasoning. Therefore, as a

[1]For example, in Piaget's theory the influence of external conditions (like task characteristics) on performance was hardly discussed.

fourth prediction it may be hypothesized that inferring a transitive relationship in a physical type-of-content task is easier than in a nonphysical type-of-content task.

- *Information Processing Theory*. Firstly, the formation of a linear ordering and the memory of the premises are expected to be easier when the premises are presented simultaneously than when they are presented successively. Secondly, because it is more difficult to form a linear ordering of a mixed-format task, it may be expected that mixed inequality-equality tasks are more difficult than equality or inequality tasks. Although information processing theorists do not use equality-format tasks to study transitive reasoning, these tasks may be expected to be easier than inequality-format tasks because the internal representation of an equality task is easier than the internal representation of an inequality task. Thirdly, according to the mixed model of Sternberg (1980b) both a verbal and a spatial ability are needed to solve linear syllogisms. For verbally presented tasks, both abilities are required, and for physical tasks, only the spatial ability is required. Thus, it may be hypothesized that verbal tasks (linear syllogisms) are more difficult than physical tasks.
- *Fuzzy Trace Theory*. Firstly, because the retrieval of a fuzzy trace is easier for simultaneously presented tasks (which contain a spatial-order correlation) than for successively presented tasks (in which the ordering of the premises is less obvious) (Brainerd & Reyna, 1992), successive presentation is expected to be more difficult than simultaneous presentation. Secondly, because it is difficult to reduce the pattern information of the mixed inequality-equality format into a fuzzy trace, it can be hypothesized that the mixed inequality-equality format is more difficult than the equality or the inequality format. Thirdly, when a fuzzy trace is used to infer the transitive relation only pattern information and no verbatim information (like type of content of tasks) is involved. Thus, different types of content are not expected to influence the difficulty level.

A summary of the influence of task characteristics on the difficulty level according to the theories is given in Table 1.

TABLE 1.
Comparison of the theories with respect to the number of abilities and influence of task characteristics on difficulty level of tasks

Theory	Topic	Predictions
Piaget	Number of abilities:	two, functional and operational reasoning
	Presentation:	successive more difficult than simultaneous
	Format:	all formats same difficulty
	Content:	verbal content more difficult than physical content
Information Processing	Number of abilities:	one (linear ordering), two (mixed model)
	Presentation:	successive more difficult than simultaneous
	Format:	equality easier than other formats, mixed more difficult than other formats
	Content:	verbal content more difficult than physical content
Fuzzy Trace	Number of abilities:	one
	Presentation:	successive more difficult than simultaneous
	Format:	equality easier than other formats, mixed more difficult than other formats
	Content:	physical content and verbal content equally difficult

1.2.5. Responses

Cognitive theories not only disagree about the kinds of tasks that should be used to measure transitive reasoning, but also about the types of responses that are required to verify that a child had really drawn a transitive inference. Piaget asked children to verbally explain their answers to verify whether a child has really used operational reasoning to solve a transitive reasoning task. According to Piaget, children were capable of operational reasoning when they could mention aloud all the premises involved (Piaget, 1961; Piaget & Inhelder, 1941; Piaget et al., 1948). More recently, Chapman & Lindenberger (1992) assumed a child to be able to draw a transitive inference when (s)he was able to explain the judgments. However, information processing theory hypothesized that the verbal explanations interfered with the cognitive processes (see, e.g., Brainerd, 1977). Also, the internal representation was not assumed to be necessarily verbal. Instead, cognitive processes were measured using reaction times (e.g., Trabasso et al., 1975) or using the performance of children on specific task formats (e.g., Murray & Youniss, 1968; Smedslund, 1963).

When the aim of a study is to construct a transitive reasoning task for determining the age of emergence as exactly as possible, using either the judgment or the judgment-plus-explanation may highly influence the result. For example, although a fair comparison between studies using different task formats could not be made, Bryant and Trabasso (1971) found children of only four years old to be able of transitive reasoning, but Chapman & Lindenberger (1992) did not find children able of transitive reasoning before the age of seven.

In fact, the discrepancy of judgment and judgment-plus-explanation approaches can be summarized as a choice between type I and type II errors (Smedslund, 1969). Given the null hypothesis that children do not have a transitive reasoning ability, a judgment-only response is prone to evoke a type I error (false positive), assuming that a child is able to draw a transitive inference when in fact it is not. However, when a verbal explanation is required, a type II error (false negative) is likely to occur, by assuming that a child is not able to draw a transitive inference when in fact it is. This inference may be caused by the child's underdeveloped verbal ability. When the aim of the study is to obtain an impression of the processes involved in the development of transitive reasoning, the explanations given by the child are useful, accepting the risk of a type II error and being somewhat conservative about the age of emergence. Using judgment-plus-explanation data, Verweij, Sijtsma, and Koops (1999) showed that several transitive and nontransitive strategies were used to solve different kinds of transitive reasoning tasks. For several task types, different strategies led to correct answers.

1.3. Goal of Present Study

The disagreement about the number of abilities involved in transitive reasoning, the type of responses to be recorded, and the influence of task characteristics on task performance led to three hypotheses:

1. H_0: Two qualitatively different abilities, functional and operational reasoning, explain the response patterns on various tasks containing transitive relations.
   H_A: One ability explains the response patterns on various transitive reasoning tasks. The tasks differ only in difficulty.
2. H_0: The response patterns based on strategy scores provide a better scale than the response patterns based on product scores (see section 2.6 for a description of strategy and product scores).
   H_A: Response patterns containing strategy scores and response patterns containing product scores both provide good scales.

3. H_0: The difficulty of transitive reasoning tasks is not influenced by task characteristics or combinations of task characteristics.
   H_A: The difficulty of transitive reasoning tasks is influenced by task characteristics or combinations of task characteristics.

For determining the number of abilities involved in transitive reasoning (first hypothesis), nonparametric item response theory (NIRT) methods (Molenaar & Sijtsma, 2000; Stout, 1993, 1996) were used to investigate the underlying dimensionality of a data set generated by means of a set of tasks having different characteristics. When one ability is involved, the task scores can be explained by one underlying dimension. Then, the transitive reasoning tasks differ only in difficulty as predicted by linear ordering theory (Trabasso et al., 1975) and fuzzy trace theory. When two or more abilities are involved for solving different kinds of tasks, multiple dimensions are needed to describe the responses of children to a set of transitive reasoning tasks.

To investigate which kind of response information gives the most useful insights into transitive reasoning, two kinds of responses were compared (second hypothesis). First, we collected the correct/incorrect judgments children gave on a set of transitive reasoning tasks (quantified as *product scores*). Second, the verbal explanations children gave for the judgments (quantified as *strategy scores*) were recorded. Before comparing the usefulness of both types of responses, the relationship between the two types was investigated. IRT models were used to compare the quality of the product scores and the strategy scores.

The predictions of the theories with respect to the difficulty level of transitive reasoning tasks (Table 1) were studied by determining the influence of task characteristics on the difficulty level of the tasks (third hypothesis). For this purpose a multiple regression model was used.

2. Method

2.1. Operationalization of the Construct

For constructing transitive reasoning tasks, three kinds of task characteristics were used. The first characteristic was presentation form of the premises. According to Piaget's theory, qualitatively different reasoning abilities are involved in successive or simultaneous presentation of the premises, while information-processing theory and fuzzy trace theory assume that one ability is involved in both presentation forms. The second characteristic was task format. Various task formats may have a different influence on the formation of a linear ordering or the use of logical rules. The third characteristic was task content. This characteristic was chosen to evaluate the influence of different kinds of content of the transitive relation on performance. According to Sternberg (1980a, 1980b), both a spatial and a verbal representation are involved in solving tasks having a verbal content (linear syllogism) whereas only a spatial representation is involved when the content is physical. The performances on the tasks were both evaluated by means of the correct/incorrect answers and the verbal explanations of the answers.

2.2. Tasks

Three kinds of task characteristics, *presentation form*, *task format*, and *task content*, with 2, 4, and 2 levels, respectively, were completely crossed, forming $2 \times 4 \times 2 = 16$ tasks. The task characteristics and their levels are:

- *Presentation form*. The two levels are:

  1. *Simultaneous presentation* (Figure 1, tasks 1, 4, 5, 7, 10, 11, 13, and 16). When the premises were presented simultaneously, all the objects were visible simultaneously dur-

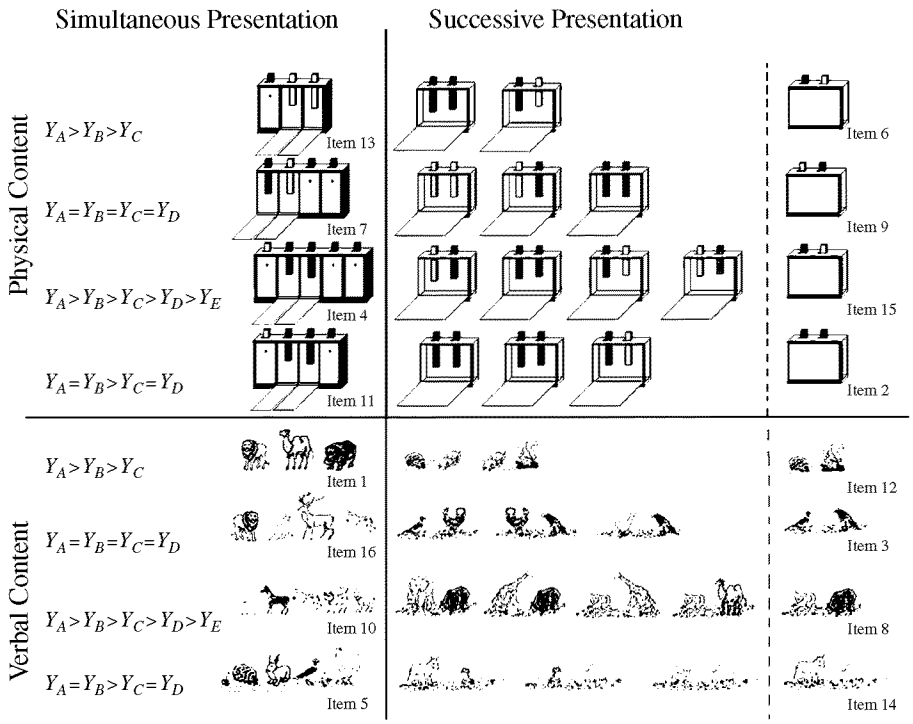

FIGURE 1.
Items of the transitive reasoning test. In the physical content items sticks had different colors (not visible here).

ing the whole task. According to Piaget's theory, this kind of task may be solved using functional reasoning.

2. *Successive presentation* (Figure 1, tasks 2, 3, 6, 8, 9, 12, 14, and 15). When the premises were presented successively, in each step of the presentation one pair of objects was visible but the other objects used in the task were not. According to Piaget's theory, this kind of task must be solved using operational reasoning.

- *Task format*. The four levels are:

1. $Y_A > Y_B > Y_C$; transitive test pair Y_A, Y_C (Figure 1, tasks 1, 6, 12, and 13). In Figure 1, Task 1, the lion is assumed to be older than the camel, and the camel is assumed to be older than the hippo.
2. $Y_A = Y_B = Y_C = Y_D$; transitive test pair Y_A, Y_C (Figure 1, tasks 3, 7, 9, and 16). In Figure 1, Task 7, all sticks have the same length.
3. $Y_A > Y_B > Y_C > Y_D > Y_E$; transitive test pair Y_B, Y_D (Figure 1, tasks 4, 8, 10, and 15). In Figure 1, Task 4, the green stick is longer than the red one, the red one is longer than the purple one, the purple one is longer than the yellow one, and the yellow one is longer than the orange one.
4. $Y_A = Y_B > Y_C = Y_D$; transitive test pair Y_A, Y_C (Figure 1, tasks 2, 5, 11, and 14). In Figure 1, Task 5, the hedgehog is assumed to be the same age as the rabbit, the rabbit is assumed to be older than the duck, and the duck is assumed to be the same age as the chicken.

- *Type of content*. The two levels are:

1. *Physical content* (Figure 1, tasks 2, 4, 6, 7, 9, 11, 13, and 15). When the content of the task was physical, the length relation between the sticks could be observed visually during the presentation of the premises.
2. *Verbal content* (Figure 1, tasks 1, 3, 5, 8, 10, 12, 14, and 16). When the content of the task was verbal, the experimenter told the age relation between the animals to the child during the presentation of the premises.

2.3. Instrument

The transitive reasoning computer program `Tranred` (Bouwmeester & Aalbers, 2002) was an individual test, constructed especially for this study. This computer program replaced the normally used *in vivo* presentation of the tasks. The advantage of a computerized test was that the administration of the test was highly standardized. Moreover, movements and sounds could be implemented to enhance the test's attractiveness and hold the child's attention. Finally, the registration of the test scores was done mostly by the program during the test administration. The verbal explanation the child gave after (s)he had clicked on the preferred answer was recorded in writing by the experimenter. The tasks were presented in the same fixed order for every subject (see Figure 1 for the task ordering). Relatively difficult tasks were alternated with easier tasks to keep the children motivated. A pilot study showed that the verbal explanations with respect to the same objects appearing in different tasks were hardly ever confused. Nevertheless, to avoid a dependence between the objects of different tasks, tasks sharing the same objects or task characteristics were alternated as much as possible by tasks having different objects or task characteristics.

2.4. Procedure

The test was administrated in a quiet room in the school building. The experimenter started a little conversation with the child to put him/her at ease and introduce the task types. Then the child did some exercises to get used to the `Tranred` program. The buttons of the program were explained. It was explained that the colored sticks could have different lengths, which could only be observed when the doors of the box were opened (see Figure 1, physical content). Also, it was explained that the animals could have different ages, but that this was not observable. After the instructions were given, the test was started.

When the content of the relation was physical, a box appeared on the screen which either contained all objects (Figure 1, *simultaneous presentation of physical content*) or a pair of objects (Figure 1, *successive presentation of physical content*). The doors were opened to show the objects of the first premise pair, and the child was asked which stick was longer or whether the sticks had the same length. When the sticks differed in length, the difference could be observed clearly. Then the child clicked on the longest stick, or on the equality button when both sticks had the same length. The doors closed and the doors of the next premise pair opened. The question was repeated for all premise pairs. During the test phase, the doors were closed and the length of the sticks could not be compared visually. The child was asked which of two sticks was longer or whether the sticks had the same length. After the child had clicked on one of the sticks or on the equality button, (s)he was asked to explain the answer. The experimenter wrote down the explanation, the box disappeared from the screen, and the next task started.

When the content of the relation was verbal, all animals (Figure 1, *simultaneous presentation of verbal content*) or a pair of animals (Figure 1, *successive presentation of verbal content*) walked onto the screen. For each premise pair, the experimenter told the child which animal was

older or that both animals had the same age. The child was asked to click on the oldest animal or on the equality button when both animals had the same age. This was repeated for all premise pairs. In the test phase, the child was asked which of two animals was older or whether both animals had the same age. After the child had clicked on one of the animals or on the equality button, the experimenter asked the child for an explanation of the answer. The experimenter wrote down the explanation, the animals walked off the screen, and the next task started.

The administration of the test took about half an hour, depending on the age of the child. For young children the test took more time and for older children the test took less time.

2.5. *Sample*

The transitive reasoning test was administered to 615 children ranging in age from 6 to 13 years old. Children came from six elementary schools in the Netherlands. The children came from middle-class social-economic status (SES) families. Table 2 gives an overview of the number of children and their mean age within each grade.

2.6. *Responses*

Product Scores When children clicked on the correct object in the test phase, they received a score of 1. When they clicked on an incorrect object a score of 0 was registered.

Strategy Scores This study builds on previous research on scaling transitive reasoning by Verweij (1994). He found satisfactory inter-rater agreement for two raters who independently coded the verbal explanations given by children who solved transitive reasoning tasks. Figure 2 gives an overview of the transitive and nontransitive strategies children used in this study to solve the 16 tasks. When children did not give an explanation they said that they had either guessed, did not know how they knew the answer, or could not explain their answer. When children gave an explanation but the premise information was not used, children used external information instead to explain their answer (e.g., *the parrot is older because parrots can live more than 40 years*); or they used visual aspects of the task to explain their answer (e.g., *the blue stick is longer because I can see that when I look close*).

When the information of the premises was used correctly, children literally mentioned the premises or reduced the information of the premises. When the premises were mentioned literally, the child mentioned all the premises involved (e.g., $Y_A > Y_B > Y_C$*: animal A is older than animal C because animal A is older than animal B, and animal B is older than animal C*). This

TABLE 2.
Number of children, mean age, and standard deviation (SD) by grade

		Age	
Grade	Number	Mean[a]	SD
2	108	95.48	7.81
3	119	108.48	5.53
4	122	119.13	5.37
5	143	132.81	5.17
6	123	144.95	5.34

[a]Number of months.

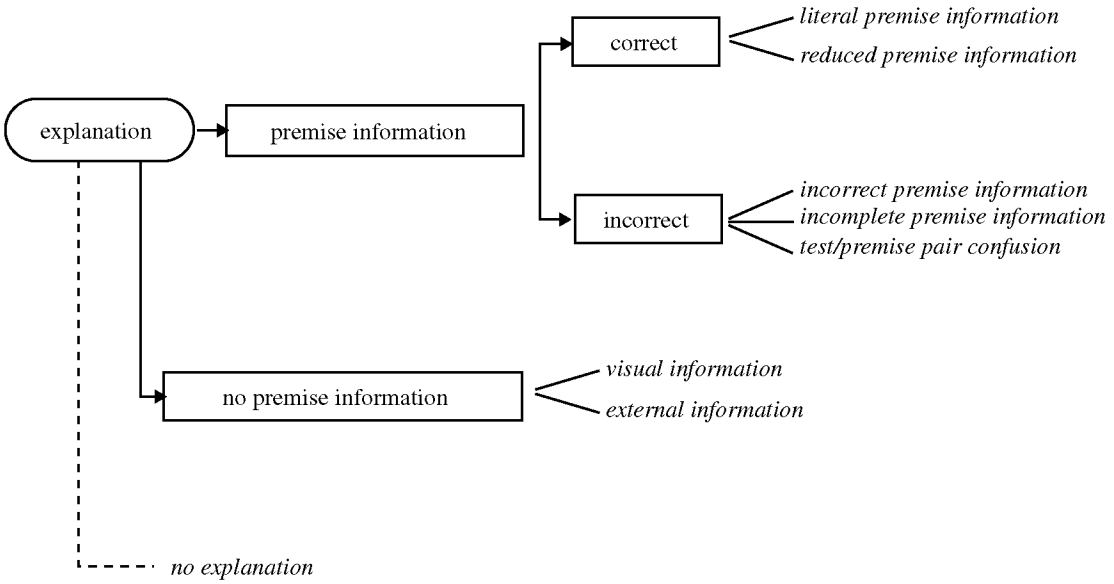

FIGURE 2.
Transitive and nontransitive reasoning strategies.

strategy is equivalent to operational reasoning in Piaget's theory. When the information of the premises was reduced correctly, children used a reduction of the premise information, by using the position of the objects (e.g., $Y_A > Y_B > Y_C > Y_D > Y_E$, simultaneous presentation; *all animals are ordered from left to right, the oldest animal first, so animal B is older than animal D*); the time sequence (e.g., $Y_A > Y_B > Y_C > Y_D > Y_E$, successive presentation; *the sticks are ordered in time, stick A was presented first and is the longest. Object B was presented before object D, so object B is longer*); or a total reduction (e.g., $Y_A = Y_B = Y_C = Y_D$: *all animals have the same age*). When the premises were mentioned incorrectly, children used an incorrect interpretation of the premises (e.g., $Y_A = Y_B > Y_C = Y_D$: *all sticks are equally long, except for stick B, which is longer. So stick A and stick C are equally long*); gave an incomplete explanation (e.g., $Y_A > Y_B > Y_C$: *stick A is longer than stick C because stick B is longer than stick C*); or confused the test-pair with a premise-pair (e.g., $Y_A > Y_B > Y_C$: *stick A is longer than stick C because I have just seen that stick A is longer than stick C*).[2] The strategies in which the premise information was mentioned literally or reduced correctly, were called *transitive reasoning strategies* and received a score of 1. All other strategies received a score of 0. In 0.16% of all cases, the explanation given by the child could not be classified in one of the strategy groups. In those cases a missing value was registered.

2.7. Item Response Theory

Our three hypotheses were investigated by means of IRT. Figure 3 gives an overview of the successive steps that were followed in this study. We first mention these steps and provide a global description of the rationale behind them. Then we explain the assumptions, methods and models in some detail.

[2]In a study by Bouwmeester, Sijtsma, and Vermunt (2004), a nominal variable was used in which all strategies were distinguished to determine the relationships between age, strategy use, and task characteristics.

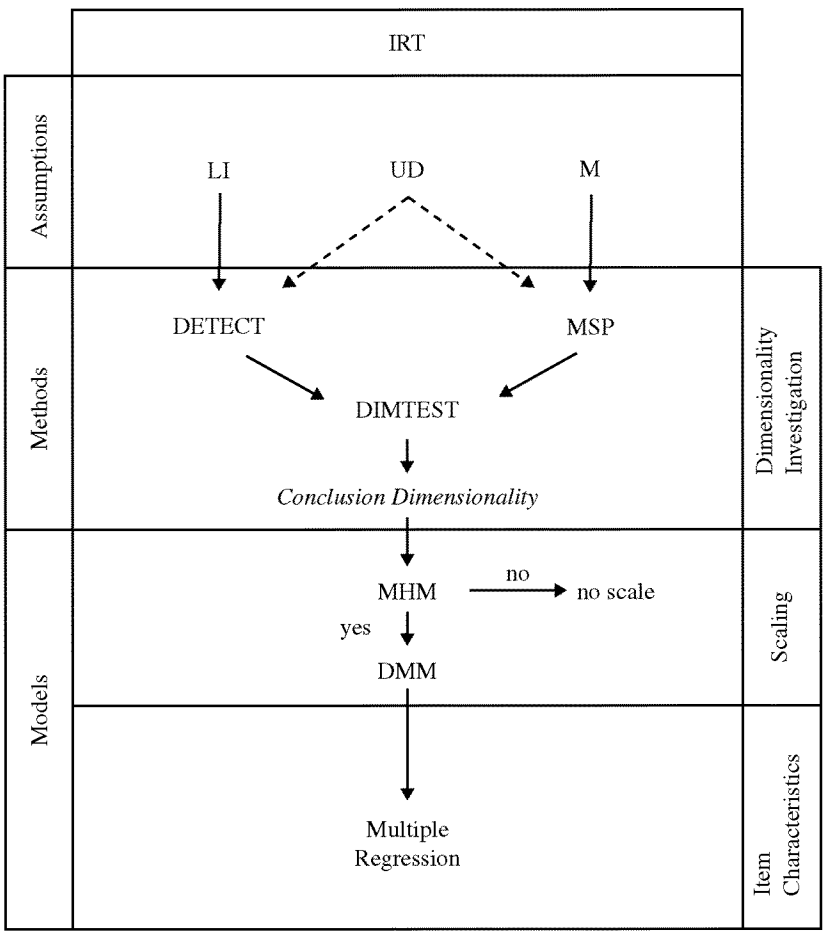

FIGURE 3.
Overview of the successive analyses.

IRT models provide methods to assess the dimensionality of the data, and thus can be used to determine the number of abilities involved in our transitive reasoning test. The program DETECT (Stout, 1996), was used to investigate dimensionality using the *local independence* assumption of IRT, and the program MSP (Molenaar & Sijtsma, 2000) was used for the same purpose using the *monotonicity* assumption of IRT. DETECT and MSP are exploratory methods. In contrast, the program Improved DIMTEST (Stout, 1993) was used to test the hypotheses about the dimensionality resulting from DETECT, MSP, and the theories about transitive reasoning.

Our approach is more exploratory than confirmatory, and there is a methodological and a theoretical reason for this. Methodologically, the exploratory methods DETECT and MSP were used instead of a confirmatory method like factor analysis, because factor analysis of dichotomous item scores has problems due to the extreme discreteness of such scores (Nandakumar, Yu, Li, & Stout, 1998; McDonald, 1985; Hatti, Krakowski, Rogers, & Swaminathan, 1996). Van Abswoude, Van der Ark, and Sijtsma (2004) argued that DETECT and MSP do not suffer from these problems. Theoretically, we chose an explorative approach because Piaget's theory is not explicit about the role of task characteristics with respect to the kind of ability (functional or operational) that is involved in transitive reasoning; that is, precise hypotheses about the task loadings on different factors or dimensions cannot be posited. However, some less explicit expectations may

be derived from the literature. Improved DIMTEST was used to test the expectation that successive tasks are solved by operational reasoning while simultaneous tasks are solved by functional reasoning (Chapman & Lindenberger, 1988, 1992).

The results of MSP, DETECT, and Improved DIMTEST were compared and the resulting conclusion answered the first hypothesis about the number of abilities. This conclusion was used as the input for investigating the second hypothesis. This was done by fitting two progressively more restrictive IRT models to the data. First, we fitted the nonparametric monotone homogeneity model (MHM; Mokken, 1971, chap. 4; Sijtsma & Molenaar, 2002, chap. 2) to the two data sets. This model implies the ordering of children with respect to ability level. A more restrictive nonparametric model is the double monotonicity model (DMM; Mokken, 1971, chap. 4; Sijtsma & Molenaar, 2002, chap. 6). When this model fits, both the children and the transitive reasoning tasks can be ordered, but on separate scales. The linear logistic test model (LLTM; Fischer, 1973, 1995; Scheiblechner, 1972) can be used to model the relationships between task difficulty and task characteristics. However, since the LLTM is a specialization of the Rasch model it is highly restrictive. Because the Rasch model did not fit our data, multiple regression on P-values was used as an alternative (Green & Smith, 1987).

2.7.1. Assumptions Common to the IRT Models Used in This Study

Local Independence Let the test consist of J dichotomously scored tasks, and let θ denote the latent ability measured by the J tasks. If the tasks measure more than one ability, we assume W latent ability parameters collected in a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$. Let X_j be the random variable for the score on task j, with $j = 1, \ldots, J$; and let x_j be the realization of this variable, with $x_j = 0, 1$. The task score variables are collected in $\mathbf{X} = (X_1, \ldots, X_J)$, and the realizations in $\mathbf{x} = (x_1, \ldots, x_J)$. Finally, the conditional probability of a 1 score on task j is denoted $P_j(\boldsymbol{\theta})$; this is the item response surface. For scalar θ, $P_j(\theta)$ is the item response function (IRF). The assumption of local independence (LI) is defined as

$$P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} P_j(\boldsymbol{\theta})^{x_j}[1 - P_j(\boldsymbol{\theta})]^{1-x_j}. \tag{1}$$

LI means that a subject's response to a task is not influenced by his/her responses to the other tasks in the test. LI implies that the covariance of two tasks, j and k, given the latent trait composite, $\boldsymbol{\theta}$, is zero; that is, $\mathrm{Cov}(X_j, X_k \mid \boldsymbol{\theta}) = 0$. This zero conditional covariance is known as weak local independence, which is important for practical item selection (Stout et al., 1996; Zhang & Stout, 1999a) to be discussed shortly.

Unidimensionality The assumption of unidimensionality (UD) means that the data structure can be explained by a unidimensional latent trait, θ. When UD does not hold, one ability is not enough to explain the variation in the scores on different tasks, and a second ability may be necessary to explain the variability, and perhaps a third, a fourth, and so on. Although UD and LI are mathematically not the same, in practice, the same methods are used to evaluate these assumptions.

Monotonicity For unidimensional θ, we assume that the IRFs are monotone nondecreasing functions. That is, for two arbitrarily chosen fixed values of θ, say, θ_a and θ_b, we have that

$$P_j(\theta_a) \leq P_j(\theta_b), \quad \text{whenever } \theta_a < \theta_b;\ j = 1, \ldots, J. \tag{2}$$

This is the monotonicity (M) assumption. Assumption M also gives information about the dimensionality of the task set, based on the variation in the slopes of the IRFs. Suppose that the task

set is multidimensional in the sense that some tasks measure θ_1 and others measure θ_2. Because the slope of an IRF expresses the strength of the relationship of a task with the latent ability or a latent ability composite, it may well be that tasks measuring one ability have steeper IRFs than tasks measuring another ability. Even if a unidimensional IRT model is incorrectly hypothesized for these multidimensional data, the slopes of the IRFs may provide evidence of this multidimensionality (Hemker, Sijtsma, & Molenaar, 1995; Mokken, 1971; Sijtsma & Molenaar, 2002, chap. 5; Van Abswoude et al., 2004). In this study, we investigated whether all the tasks measure the same θ and, in case of multidimensionality, we tried to identify unidimensional subsets of tasks.

The Monotone Homogeneity Model The MHM (Mokken, 1971, chap. 4; Sijtsma & Molenaar, 2002, chap. 2) is based on the assumptions of LI, UD, and M. The MHM is an NIRT model that orders subjects on the θ scale using their number-correct score, defined as $X_+ = \sum X_j$ (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997). Theoretically, this ordering of persons is the same for each task, and also for a number-correct score, based on the task scores Y_j from any subset of tasks selected from the larger set of tasks that are driven by θ and agree with the MHM. In practice, the number of tasks affects the accuracy of a person ordering estimated by means of the number-correct score X_+.

2.7.2. *Methods to Assess the Dimensionality of the Data*

We used three methods to assess the dimensionality structure of the two dichotomous data sets. The first method was the item selection procedure in the computer program MSP (Molenaar & Sijtsma, 2000; Sijtsma & Molenaar, 2002, chap. 5). This procedure is used to select the tasks on the basis of assumption M. The second item selection method was DETECT (Zhang & Stout, 1999b). The third method was Improved DIMTEST (Stout, Froelich, & Gao, 2001). This method was used to test the null-hypothesis of UD for the whole task set. Both DETECT and DIMTEST use the assumption of LI to assess UD.

Program MSP MSP (Molenaar & Sijtsma, 2000) uses scalability coefficient H (Mokken, 1971, pp. 157–169) to assess the discrimination power of individual tasks (i.e., the slopes of the IRFs) and the whole test. The item coefficient H_j is an index of the slope of the IRF relative to the spread of the number-correct score X_+ in the group under consideration. The higher H_j, the better task j discriminates between different X_+ scores. The H coefficient for the whole test of J tasks summarizes the slope information contained in all J item coefficients H_j.

Mokken, Lewis, and Sijtsma (1986) argued that higher positive H values reflect higher discrimination of the whole set of tasks and, thus, a more accurate ordering of subjects. In practical test construction, to have at least reasonable discrimination, a lower bound value for H_j and H of 0.3 is recommended (Mokken, 1971, p. 184). Other guidelines (Sijtsma & Molenaar, 2002, p. 60) for the interpretation of H are: $0.3 \leq H < 0.4$ is a weak scale; $0.4 \leq H < 0.5$ is a medium scale; and $0.5 \leq H < 1.0$ is a strong scale. The MSP item selection procedure has been described in detail by Mokken (1971, pp. 190–194; also see Molenaar & Sijtsma, 2000; Sijtsma & Molenaar, 2002, chap. 5). It is a bottom-up procedure, that starts by selecting the two items with the highest significantly positive H_{jk} that is at least c ($c > 0$; user-specified). Then the procedure adds tasks one by one, in each step maximizing the total H of the selected items, such that $H_j \geq c$ for all selected items (for possible exceptions, see Sijtsma & Molenaar, 2002, p. 79). After having selected the first scale, the procedure continues by selecting from the unselected items a second scale, a third scale, and so on. Van Abswoude et al. (2004) found that MSP was able to exactly retrieve the true dimensionality from simulated data when latent traits did not correlate highly (say, higher than .4). Hemker et al. (1995; see also Sijtsma & Molenaar, 2002,

p. 81; Van Abswoude et al., 2004) recommended using a range of c values from $c = 0.00$ to $c = 0.55$ with increments of 0.05, and described sequences of outcomes for increasing c values typical of multidimensionality and unidimensionality.

Program DETECT The computer program DETECT (Zhang & Stout, 1999a, 1999b; Roussos, Stout, & Marden, 1998) contains an item selection algorithm that tries to find the partitioning $\mathcal{P}$ for which the degree to which LI is satisfied is maximal, given all possible partitions of the task set. In contrast to MSP, where assumption M is the basis of the item selection, weak LI is the basis of DETECT. DETECT works best when all individual tasks load on one θ (but not necessarily the same θ for all tasks). This is called approximate simple structure (Zhang & Stout, 1999a). When individual tasks load on different θs, approximate simple structure does not hold and no best partitioning can be determined. Under the assumption of approximate simple structure, the DETECT index is maximal when the underlying structure is correctly represented by the number and the composition of the clusters. When the DETECT value is zero, no best partitioning is possible and the task set is unidimensional. As a rule of thumb (Zhang & Stout, 1999b), a task set is considered unidimensional when the DETECT value is smaller than 0.1. To evaluate whether approximate simple structure exists, Zhang and Stout (1999b) proposed that their index $R \geq 0.8$. When approximate simple structure does not exist, it is difficult to decide how many dimensions are involved. Van Abswoude et al. (2004) recommended using MSP and DETECT together for analyzing one's data.

Program Improved DIMTEST DIMTEST is a procedure that tests the null hypothesis that a set of items is dimensionally similar to another set of items. Because the DIMTEST procedure does not work for short tests, we used the improved DIMTEST procedure (Nandakumar & Stout, 1993). This procedure generates a unidimensional data set using a nonparametric bootstrap method to correct for bias in parameter estimates and to increase the power of the DIMTEST statistic (Stout et al., 1995). The hypothesis is tested that the generated data set has the same dimensionality as the real data set. For example, we tested the hypothesis that the responses to the successively presented tasks are dimensionally distinct from those to the simultaneously presented tasks. We considered the simultaneously presented tasks to be the Assessment Test (AT; see Nandakumar & Stout, 1993) and the successively presented tasks to be the Partition Test (PT; see Nandakumar & Stout, 1993). The items in AT were hypothesized to measure one dominant trait. An asymptotic test statistic, denoted T, was used to test whether the items in AT and PT measure the same θ.

2.7.3. *IRT Models and Assessment of Fit*

Monotone Homogeneity Model After the dimensionality of the transitive reasoning data was investigated, the computer program MSP (Molenaar & Sijtsma, 2000) was used to investigate the fit of the MHM to the two data sets. To evaluate whether the IRFs of the J tasks were all nondecreasing, subjects were partitioned into J restscore groups on the basis of their restscore, $R_{(-j)} = X_+ - X_j$. The restscore $R_{(-j)}$ is an ordinal estimator of θ (Junker, 1993). To enhance power, small adjacent restscore groups were joined using recommendations given by Molenaar and Sijtsma (2000, p. 100). For each restscore group r the probability of giving a correct answer, $P(X_j = 1 \mid R_{(-j)} = r)$, was estimated, and the hypothesis was tested that these probabilities are nondecreasing in $R_{(-j)}$.

Double Monotonicity Model The DMM adds a fourth assumption to the MHM, which states that the IRFs do not intersect. This fourth assumption equals *invariant item ordering*; that is,

the ordering of the J tasks is the same for different subgroups of subjects (except for possible ties), including individual θs. In particular, for two tasks j and k, if we know for one θ_0 that $P_j(\theta_0) < P_k(\theta_0)$, then it follows that for any θ, we have that $P_j(\theta) \leq P_k(\theta)$. This ordering property can be extended to all J tasks simultaneously.

MSP was used to investigate whether the IRFs intersected. The scalability coefficient H^T (Sijtsma & Meijer, 1992) for the J tasks in the test and the person coefficient H_i^T were used to evaluate intersection of the IRFs. As a rule of thumb, if $H^T \geq 0.3$ and the percentage of negative H_i^T values < 10, then the IRFs do not intersect. Three additional methods were used to investigate the nonintersection of IRFs for pairs of tasks. These methods are the *restscore method*, the *restsplit method*, and the inspection of the *P-matrices*, $P(-, -)$ and $P(+, +)$ (Sijtsma & Molenaar, 2002, chap. 6). These methods are based on the same rationale, but use different subgroupings of respondents for estimating the IRFs. The three methods differ in accuracy to estimate the IRFs and in power to detect intersections.

Linear Regression Using P-values In the multiple regression model the proportions correct are regressed on the task characteristics. Because the proportions corrected are bounded between 0 and 1, a logistic transformation of the P-values was used.

3. Results

3.1. Relation between Product Scores and Strategy Scores

Table 3 shows the proportions of strategy use and the proportions of correct answers given strategy use. The two "correct" strategies (literal and reduced premise information) almost always led to correct answers. The three strategies in which no premise information is used (visual information, external information, and no explanation) have proportions of correct answers close to chance level. Test/premise pair confusion relatively often led to a correct answer, although it is an incorrect strategy. Table 3 shows that incorrect strategies often led to correct answers that were produced by chance.

3.2. Hypothesis 1: Assessing Dimensionality

3.2.1. Analysis of Product Scores

Twelve cases were rejected from the analysis because of missing values on one or more tasks. The resulting sample consisted of 603 subjects.

TABLE 3.
Strategy use and proportion of correct answers

Strategy	Proportion of strategy use	Proportion of correct answers
Literal complete premise information	0.16	0.94
Reduced premise information	0.21	0.97
Incorrect premise information	0.19	0.23
Incomplete premise information	0.10	0.48
Test/premise pair confusion	0.10	0.58
Visual information	0.06	0.35
External information	0.03	0.36
No explanation	0.16	0.37

TABLE 4.
Item selection for increasing c-values, for MSP analysis using product scores

c	Scale 1	Scale 2	Scale 3	Scale 4	# Tasks rejected
0.00	1,3,4,7,9,16	5,6,8,10,11,12,13,14,15			1
0.05	1,3,4,7,9,16	5,6,8,10,11,12,13,14,15			1
0.10	1,3,4,7,9,16	5,6,8,10,11,12,13,14,15			1
0.15	3,4,7,9,16	1,5,8,10,11,12,13,14,15			2
0.20	7,9,16	1,4,5,8,10,11,12,13,14,15			3
0.25	7,9,16	5,14,10,8,1	4,13	11,15	4
0.30	7,9,16	5,10,8,1	4,13		7
0.35	7,9,16	5,10,1	4,13		8
0.40	9,16	10,1	4,13		10
0.45	9,16	10,1	4,13		10
0.50	9,16	10,1	4,13		12
0.55					16

MSP Analysis Table 4 shows the sequence of outcomes of the MSP analysis with increasing c-values. Task 2 was immediately rejected because of a negative covariance with one of the other tasks. For lowerbound $c = 0$, two scales were formed containing six and nine tasks, respectively, which suggests that the test measures at least two latent abilities. For increasing c-values, Task 3 and Task 6 were also rejected, and a third and a fourth scale were formed, both containing two tasks. For c-values of 0.40 and higher, almost all tasks were rejected and no scale was formed containing more than two tasks. For $c = 0.55$ no scale was formed. On the basis of the guidelines of Hemker et al. (1995), it was concluded that at least two abilities were involved in answering the tasks. One scale contained the tasks 7, 9, and 16 ($H = 0.44$), which all have the format $Y_A = Y_B = Y_C = Y_D$, and another, rather weak ($H = 0.25$) scale contained the tasks 1, 4, 5, 8, 10, 11, 12, 13, 14, and 15, which have the formats $Y_A > Y_B > Y_C$; $Y_A > Y_B > Y_C > Y_D > Y_E$; and $Y_A = Y_B > Y_C = Y_D$.

DETECT Analysis A random half of the sample was used for the DETECT procedure. The second half of the sample was used for cross-validation. The R index for assessing simple structure was 0.74. This is smaller than the value of at least 0.8 that Zhang and Stout (1999b) proposed for approximate simple structure; refer to this source for a discussion on how to deal with situations like this one. The maximum DETECT value [denoted $D_\alpha(\mathcal{P}^*)$] was 0.88, which was higher than 0.1, indicating that the task set was not unidimensional. The partitioning with this value had three clusters. For the second half of the sample, using the same partitioning that was found to be optimal for the first data set, we found $D_\alpha(\mathcal{P}^*) = 0.48$ and $R = 0.43$. To gain more insight into the dimensionality of the data, 20 random samples of approximately 50% of the subjects were drawn from the original sample and the DETECT value was calculated for each sample. Figure 4 shows the number of times that two tasks were in the same cluster. Three (overlapping) clusters can be distinguished. One contained the tasks 3, 7, 9, and 16 (all with format $Y_A = Y_B = Y_C = Y_D$), which were almost always in the same cluster. A second cluster contained the tasks 1, 5, 8, 10, 11, and 14, and a third cluster contained the tasks 2, 4, 12, and 13. Task 6 did not fit well in any of the clusters and Task 15 might belong to either the second or the third cluster.

Improved DIMTEST Analysis Three hypotheses were tested. First, it was tested whether the tasks that were simultaneously presented measured the same ability as the tasks that were

	3	**7**	**9**	**16**	**6**	**1**	**5**	**8**	**10**	**11**	**14**	**15**	**2**	**4**	**12**	**13**
3		19	19	19	9	4	0	0	0	0	0	1	1	6	0	4
7	19		20	20	10	4	0	0	0	0	0	0	0	5	0	3
9	19	20		20	10	4	0	0	0	0	0	0	0	5	0	3
16	19	20	20		10	4	0	0	0	0	0	0	0	5	0	2
6	9	10	10	10		6	5	5	5	3	3	2	0	0	3	0
1	4	4	4	4	6		15	16	16	11	9	2	0	1	2	1
5	0	0	0	0	5	15		20	20	15	13	5	1	0	2	0
8	0	0	0	0	5	16	20		20	15	13	5	1	0	2	0
10	0	0	0	0	5	16	20	20		15	12	5	1	0	2	0
11	0	0	0	0	3	11	15	15	15		14	9	2	0	0	0
14	0	0	0	0	3	9	13	13	12	14		11	5	2	4	2
15	1	0	0	0	2	2	5	5	5	9	11		11	7	8	9
2	1	0	0	0	0	0	1	1	1	2	5	11		12	13	9
4	6	5	5	5	0	1	0	0	0	0	2	7	12		13	17
12	0	0	0	0	3	2	2	2	2	0	4	8	13	13		15
13	4	3	3	2	0	1	0	0	0	0	2	9	9	17	15	

16 through 20 times in the same cluster
10 through 15 times in the same cluster
6 through 9 times in the same cluster

FIGURE 4.
DETECT Partitioning in clusters for 20 random samples, product scores.

successively presented (Piaget's theory). Second, it was tested whether the tasks that had a verbal content measured the same ability as the tasks that had a physical content (Sternberg's mixed model). Third, it was tested whether the tasks with an equality format ($Y_A = Y_B = Y_C = Y_D$) measured another ability than the other tasks, which was the result of MSP and DETECT. The results were as follows.

- *Hypothesis 1*: Statistic T was 1.24 ($p > 0.05$), so we cannot conclude that simultaneously and successively presented tasks require different abilities.
- *Hypothesis 2*: Statistic T was 2.51 ($p < 0.05$), so the tasks having a verbal content may measure a different ability than the tasks having a physical content.
- *Hypothesis 3*: Statistic T was 2.85 ($p < 0.05$), so the equality tasks may measure a different ability than the the tasks having a inequality or mixed inequality/equality format.

Conclusion about Dimensionality of Product Scores MSP, DETECT and improved DIMTEST results converged to the conclusion that the structure of the *product scores* is not unidimensional. MSP distinguished at least two dimensions, one defined by tasks with the equality format and the other by the other tasks. DETECT found three partly overlapping clusters, one of which contained the tasks having the equality format. The Improved DIMTEST procedure supported the hypothesis that the tasks having an equality format were dimensionally distinct from the other tasks, and that the tasks having a verbal content were dimensionally distinct from the tasks having a physical content. None of the three methods showed that the successively and simultaneously presented tasks were dimensionally distinct.

3.2.2. Analysis of Strategy Scores

Fifteen subjects were rejected from the analysis because of missing values on one or more tasks. The resulting sample consisted of 600 subjects. Because only six children gave a transitive reasoning explanation for Task 2, this task was rejected from further analysis.

MSP Analysis Table 5 shows the sequence of item selection outcomes with increasing c-values. For $c = 0$, all tasks were selected into the same scale. For higher c-values, all tasks were selected into the same scale until a c-value of 0.40, when Task 12 was rejected from the scale. For $c = 0.45$, a second scale was formed containing the tasks 3, 9, and 14. Considering this sequence of outcomes, it could be concluded that the structure of the strategy scores was unidimensional.

DETECT Analysis The R ratio for the first half of the sample was 0.68, indicating that there was no approximate simple structure. The maximum DETECT value $[D_\alpha(\mathcal{P}^*)]$ was 0.57, indicating that the task set was not unidimensional. The partitioning with maximum DETECT value had two clusters. For the cross-validation sample we found that $D_\alpha(\mathcal{P}^*) = 0.24$ and $R = 0.32$. Again, 20 samples of approximately 50% of the original sample size were drawn at random from the original sample and the DETECT values were calculated for each sample. Figure 5 shows two overlapping clusters; one cluster containing the tasks 3, 7, 9, and 16, which were almost always in the same cluster, and one cluster containing the other tasks. It could not be decided to which cluster the tasks 4 and 6 belong.

Improved DIMTEST Analysis The same three hypotheses were tested as was done using the product scores. The results were as follows.

- *Hypothesis 1*: Statistic T was 0.70 ($p > 0.05$), so we could not conclude that simultaneously and successively presented tasks required different abilities.
- *Hypothesis 2*: Statistic T was 2.26 ($p < 0.05$), so the tasks having a verbal content may measure another ability than the tasks having a physical content.
- *Hypothesis 3*: Statistic T was 2.30 ($p < 0.05$), so the equality tasks may measure a different ability than tasks having an inequality or mixed inequality/equality format.

Conclusion about Dimensionality of Strategy Scores Different methods led to different conclusions about the dimensionality of the data. MSP indicated unidimensionality. Improved

TABLE 5.
Item selection for increasing c-values, for MSP analysis using strategy scores

c	Scale 1	Scale 2	# Tasks rejected
0.00	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.05	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.10	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.15	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.20	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.25	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.30	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.35	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
0.40	1,3,4,5,6,7,8,9,10,11,13,14,15,16		1
0.45	1,4,6,7,8,10,13,15,16	3,9,14	3
0.50	2,6,7,9,11,16		6
0.55	4,6,8,10,13,16	7,11	5

	1	**5**	**8**	**10**	**11**	**12**	**14**	**15**	**13**	**4**	**6**	**3**	**7**	**9**	**16**
1		15	16	19	11	13	16	14	7	6	2	1	1	1	1
5	15		15	17	14	13	20	16	6	3	4	0	0	0	0
8	16	15		17	10	16	16	16	9	7	2	0	1	0	0
10	19	17	17		12	14	18	14	8	5	3	0	0	0	0
11	11	14	10	12		6	13	13	5	3	8	3	3	3	4
12	13	13	16	14	6		13	13	12	7	6	0	0	0	0
14	16	20	16	18	13	13		13	6	3	3	0	0	0	0
15	14	16	16	14	13	13	13		11	7	4	0	0	0	0
13	7	6	9	8	5	12	6	11		16	11	2	3	3	3
4	6	3	7	5	3	7	3	7	16		12	7	8	8	8
6	2	4	2	3	8	6	3	4	11	12		8	8	9	9
3	1	0	0	0	3	0	0	0	2	7	8		19	20	19
7	1	0	1	0	3	0	0	0	3	8	8	19		19	18
9	1	0	0	0	3	0	0	0	3	8	9	20	19		19
16	1	0	0	0	4	0	0	0	3	8	9	19	18	19	

16 through 20 times in the same cluster
10 through 15 times in the same cluster
6 through 9 times in the same cluster

FIGURE 5.
DETECT Partitioning in clusters for 20 random samples, strategy scores.

DIMTEST suggested distinct abilities for both the equality tasks and tasks having a verbal content. DETECT resulted in two dimensions. One cluster contained the tasks with the equality format and the other cluster contained the other tasks. The tasks having a verbal content did not form a distinct cluster.

3.3. Hypothesis 2: Fitting the NIRT Models

The product scores did not form a unidimensional scale. Therefore, the NIRT models were only fitted to the strategy scores.

3.3.1. Analysis of Strategy Scores

MSP, DETECT, and Improved DIMTEST led to different conclusions about the dimensionality structure of the strategy scores. In particular, the equality tasks formed a distinct cluster. In the following analyses, 15 transitive reasoning tasks (except Task 2) were used.

MHM Analysis The H-value of the scale was 0.45, indicating a medium strength scale. All H_js were between 0.38 (Task 12) and 0.66 (Task 16). Table 6 gives an overview of the P_j-values and the H_j-values. The item-restscore regressions were increasing or nonsignificantly locally decreasing for each of the 15 tasks. Thus the MHM fitted the 15 tasks.

DMM Analysis The H^T value was 0.52, and the percentage of negative H_i^T values was 1.4. According to the assessment of intersection via restscore groups, the IRFs of tasks 3 and 10, and tasks 9 and 10 intersected significantly ($z_{3,10} = 1.81$; $z_{9,10} = 3.05$). Investigating the intersection via restsplit groups, the IRFs of tasks 9 and 10, and tasks 4 and 12 intersected significantly for two dichotomizations (yielding $z_{9,10}$ values of 2.04 and 3.12; and $z_{4,12}$ values of 1.66 and 1.67). The bivariate proportions in the $P(+,+)$ matrix showed an intersection of the IRFs of tasks 9 and 10.

TABLE 6.
P_j-value and H_j-value of the items, based on strategy scores

Item	Presentation	Format	Content	P_j	H_j
6	successive	$Y_A > Y_B > Y_C$	physical	.05	.46
15	successive	$Y_A > Y_B > Y_C > Y_D > Y_E$	physical	.07	.47
5	simultaneous	$Y_A = Y_B > Y_C = Y_D$	verbal	.15	.40
14	successive	$Y_A = Y_B > Y_C = Y_D$	verbal	.19	.42
8	successive	$Y_A > Y_B > Y_C > Y_D > Y_E$	verbal	.21	.48
11	simultaneous	$Y_A = Y_B > Y_C = Y_D$	physical	.31	.40
4	simultaneous	$Y_A > Y_B > Y_C > Y_D > Y_E$	physical	.39	.46
12	successive	$Y_A > Y_B > Y_C$	verbal	.40	.38
3	successive	$Y_A = Y_B = Y_C = Y_D$	verbal	.45	.41
1	simultaneous	$Y_A > Y_B > Y_C$	verbal	.56	.46
10	simultaneous	$Y_A > Y_B > Y_C > Y_D > Y_E$	verbal	.52	.51
9	successive	$Y_A = Y_B = Y_C = Y_D$	physical	.54	.40
13	simultaneous	$Y_A > Y_B > Y_C$	physical	.57	.50
7	simultaneous	$Y_A = Y_B = Y_C = Y_D$	physical	.77	.55
16	simultaneous	$Y_A = Y_B = Y_C = Y_D$	verbal	.86	.66

Summarizing the results of the four methods, the task pair (9,10) had the most serious intersections, but the violations were small. It was concluded that the DMM fitted the strategy data and that an invariant item ordering held for the 15 tasks.

3.4. Hypothesis 3: The Influence of Task Characteristics on Difficulty

3.4.1. Multiple Regression

A multiple regression analysis was performed on the 15 tasks to which the DMM fitted. The dependent variable was the logit transformation of the proportion correct of each task. The three task characteristics were the predictor variables. Because the task characteristics were nominal they were transformed to dummy variables. A significant F-value was found: $F_{6,14} = 6.77$ ($p = 0.01$). The adjusted R^2 was .71, meaning that the model explained 71% of the variance of the difficulty levels of the 15 tasks. Two regression weights (Table 7) significantly deviated from 0. The format $Y_A = Y_B = Y_C = Y_D$ had a positive effect on the easiness of a task. Simultaneous presentation was easier than successive presentation.

TABLE 7.
Estimated weights of the multiple regression model

Characteristic	B	SE	β	P-value
(Constant)	−1.980	.740		.028
$Y_A > Y_B > Y_C$	.273	.698	.096	.706
$Y_A = Y_B = Y_C = Y_D$	1.797	.698	.632	.033
$Y_A > Y_B > Y_C > Y_D > Y_E$	.221	.611	.078	.727
$Y_A = Y_B > Y_C = Y_D$	−.957	.631	−.305	.168
Presentation	1.504	.367	.597	.003
Content	.333	.393	.132	.420

Simultaneous presentation form was coded 1. Successive presentation form was coded 0. Verbal type of content was coded 1. Physical type of content was coded 0.

4. Discussion

Theories stemming from different epistemological backgrounds used different definitions, operationalizations, and methods to study transitive reasoning. This led to disagreement about the number of abilities involved in transitive reasoning, the kind of responses to be collected, and the influence of task characteristics on performance. In this study, we first evaluated the hypothesis that different abilities are involved in solving tasks by investigating the dimensionality structure of a task set with various task characteristics. Both the product scores and the strategy scores were analyzed and the results compared. Second, a scale was constructed which measured individual differences in transitive reasoning. Third, the influence of task characteristics on the difficulty level of tasks was determined.

The results of MSP, DETECT, and Improved DIMTEST for the product data and the strategy data showed that the dimensionality of successively and simultaneously presented tasks did not differ. Thus, there is no evidence that in transitive reasoning functional and operational reasoning should be distinguished. This result does not support Piaget's theory. With respect to Sternberg's mixed model, it appeared that Improved DIMTEST suggested different abilities for tasks having a verbal content and tasks having a physical content. Although MSP and DETECT did not support this finding, a tentative conclusion might be that there is some evidence that the tasks having a verbal content require an additional verbal ability. A possible explanation for finding the distinct abilities only by means of DIMTEST may be that the verbal content tasks were relatively easy linear syllogisms with respect to the verbal ability component (without negations or marked adjectives; see Sternberg, 1980b). In terms of Sternberg's mixed model, this would mean that verbal content tasks require a weak verbal component in addition to the spatial ordering component, whereas physical content tasks only require a spatial ordering component.

In contrast to the results of the past four decades of research on cognitive development (see, e.g., Brainerd, 1977; Murray & Youniss, 1968; Smedslund, 1963), we found that the strategy scores produced more straightforward and useful findings than the product scores. The data structure of the strategy scores could be explained by one dimension according to MSP, but at least three dimensions were needed to explain the data structure of the product scores. The results of the three methods did not converge to one interpretation. The multidimensionality in the product scores might best be explained by the difference in accuracy and meaning of the two types of responses. A product score of 1 means that the child had clicked on the correct object. A 1 score may therefore not represent true transitive reasoning ability, but instead may be due to additional unimportant skills or tricks. The data structure of the product scores is expected to be fuzzier than the data structure of the strategy scores, for which the meaning of a 0 or 1 score is clearer. This may explain why the product data were multidimensional and the strategy data were unidimensional.

Our population consisted of children of six years and older, which were well capable of explaining their thoughts afterwards. This population was chosen because our aim was to describe the development of transitive reasoning, but not to determine the age of emergence of transitive reasoning. This was often the aim of researchers studying transitive reasoning by young children (Braine, 1959; Bryant & Trabasso, 1971; Murray & Youniss, 1968; Smedslund, 1963). When younger children are studied, the requirement of verbal explanation may cause many false negatives due to verbal incapacity. Then, product scores are expected to be more useful.

For the strategy scores, DETECT found that the equality-format tasks ($Y_A = Y_B = Y_C = Y_D$) formed a distinct cluster. MSP and Improved DIMTEST did not find a distinct dimension for the equality-format tasks. The equality-format tasks were easy, and they discriminated well between children with low ability levels, and worse between children with higher ability levels.

Although the equality-format tasks may not be entirely dimensionally equal to the other tasks, they are useful from a practical point of view because they discriminate well at θ levels not covered by the other tasks but desirable for a transitive reasoning scale.

MSP, DETECT, and Improved DIMTEST evaluate dimensionality from different perspectives on the data. The three methods differ in several ways and each has merits and drawbacks. Van Abswoude, et al (2004) concluded that DETECT is the best method to assess true dimensionality. However, the simple structure assumption is a strong assumption which may not be realistic in many psychological settings. MSP is susceptible to locally optimal solutions because it uses a sequential clustering procedure. Further, MSP often does not accurately distinguish highly correlated abilities ($> .4$), but DETECT does. However, by forcing tasks into clusters of highly correlating traits, DETECT is vulnerable to chance capitalization. Also, Van Abswoude et al. (2004) found that DETECT does not reflect dimensionality well when abilities are measured by unequal numbers of tasks. Improved DIMTEST does not reflect true dimensionality well when abilities are measured by unequal numbers of tasks and these task subsets have equal average discrimination. DETECT and Improved DIMTEST both need large sample sizes, and Improved DIMTEST has low power for short tests. Nevertheless, when the methods are used next to each other, they can compensate each other's shortcomings and offer a detailed description of the underlying dimensionality. In future research it would be interesting to sample new data and use the results from the present study for confirmatory analysis. Multidimensional IRT models might be appropriate for this purpose (see e.g., Kelderman & Rijkes, 1994, and Reckase, 1997).

It is important to point out that statistical methods give mathematical definitions of dimensions, and that these dimensions are not equivalent to psychological abilities. The interpretation of the dimensionality of the data is dependent on the operationalization of the construct of transitive reasoning, but not directly on the construct itself. While usually no explicit distinction is made between the operationalization of the construct and the construct itself when interpreting the results, the distinction should not be ignored. In our study, we used a broad operationalization of transitive reasoning by using different kinds of task characteristics. Using this operationalization, we could explain the structure of the strategy data by means of one dimension. When we would have used a narrower operationalization based only on the theory of Piaget (e.g., see Verweij, Sijtsma & Koops, 1999), we probably would have found another dimensionality structure leading to another interpretation.

Multiple regression was used to determine the influence of task characteristics on the task difficulty level. With respect to presentation form, each of the cognitive theories predicted that simultaneous presentation was easier than successive presentation. This was indeed what was found. With respect to the task format, the equality format appeared to be easier than the other formats. This result was correctly predicted by information processing theory and fuzzy trace theory but not by Piaget's theory. Verbal and physical content hardly influenced difficulty level, and this was only predicted correctly by fuzzy trace theory.

This study showed that IRT techniques are not only useful tools to construct tests but also offer a set of methods to investigate psychological theories, in particular the dimensionality of a psychological construct. Now that we know that transitive reasoning can be explained by one dimension, further research should be done to interpret this ability in more detail. In our current research, Bouwmeester et al. (2004) used a latent class regression model, and found that several latent classes could be distinguished in which children used different patterns of correct and incorrect strategies and in which the influence of task characteristics on performance was different. From a developmental perspective, it is important to determine whether the development of the ability found in this study is continuous or discontinuous [see e.g., Hosenfield, Van der Maas, & Van den Boom (1997), and Thomas, Lohaus, & Kessler (1999), for studies on discontinuity in other Piagetian tasks]. This work is now in progress.

References

Bouwmeester, S., & Aalbers, T. (2002). *Tranred*. Tilburg: Tilburg University.

Bouwmeester, S., Sijtsma, K., & Vermunt, J.K. (2004). Latent class regression analysis to describe cognitive developmental phenomena: an application to transitive reasoning. *European Journal of Developmental Psychology, 1*, 67–86.

Braine, M. D.S. (1959). The onthogeny of certain logical operations: Piaget's formulation examined by nonverbal methods. *Monographs for the Society for Research in Child Development, 27*, 41–63.

Brainerd, C.J. (1977). Response criteria in concept development research. *Child Development, 48*, 360–366.

Brainerd, C.J., & Kingma, J. (1984). Do children have to remember to reason? A fuzzy-trace theory of transitivity development. *Developmental Review, 4*, 311–377.

Brainerd, C.J., & Kingma, J. (1985). On the independence of short-term memory and working memory in cognitive development. *Cognitive Psychology, 17*, 210–247.

Brainerd, C.J., & Reyna, V.F. (1992). The memory independence effect: What do the data show? What do the theories claim? *Developmental Review, 12*, 164–186.

Bryant, P.E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature, 232*, 456–458.

Chapman, M., & Lindenberger, U. (1988). Functions, operations, and decalage in the development of transitivity. *Developmental Psychology, 24*, 542–551.

Chapman, M., & Lindenberger, U. (1992). Transitivity judgments, memory for premises, and models of children's reasoning. *Developmental Review, 12*, 124–163.

Clark, H.H. (1969). Linguistic processes in deductive reasoning. *Journal of Educational Psychology, 76*, 387–404.

DeSoto, C.B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Social Psychology, 2*, 513–521.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fischer, G.H. (1995). The linear logistic test model. In G.H. Fischer and I.W. Molenaar (Eds.), *Rasch Models, Foundations, Recent Developments, and Applications* (pp. 131–155). New York: Springer-Verlag.

Grayson, D.A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383–392.

Green, K.E., & Smith, R.M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*, 369–381.

Hatti, J., Krakowski, K., Rogers, H.J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimmensionality. *Applied Psychological Measurement, 20*, 1–14.

Hemker, B.T., Sijtsma, K., & Molenaar, I.W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*, 337–352.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sumscore in polytomous IRT models. *Psychometrika, 62*, 331–348.

Hosenfield, B., Van der Maas, H. L.J., & van den Boom, D.C. (1997). Detecting bimodality in the analogical reasoning performance of elementary schoolchildren. *International Journal of Behavioral Development, 20*, 529–547.

Huttenlocher, J. (1968). Constructing spatial images. *Psychological Review, 75*, 550–560.

Huttenlocher, J., & Higgens, E.T. (1971). Adjectives, comparatives and syllogisms. *Psychological Review, 78*, 487–504.

Junker, B.W. (1993). Conditional association, essential independence, and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359–1378.

Kelderman, H., & Rijkes, C. P.M. (1994). Loglinear multidimensional IRT models for polytomously scores items. *Psychometrika, 59*, 149–176.

McDonald, R.P. (1985). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379–396.

Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis*. The Hague: Mouton.

Mokken, R.J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "the Mokken scale: A critical discussion." *Applied Psychological Measurement, 10*, 279–285.

Molenaar, I.W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows. A program for Mokken Scale analysis for Polytomous items [software manual]*. Groningen, The Netherlands: iecProGamma.

Murray, J.P., & Youniss, J. (1968). Achievement of inferential transitivity and its relation to serial ordering. *Child Development, 39*, 1259–1268.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*, 41–68.

Nandakumar, R., Yu, F., Li, H.H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement, 22*, 99–115.

Piaget, J. (1961). *Les Méchanicismes Perceptives*. Paris: Presses Universitaires de France.

Piaget, J., & Inhelder, B. (1941). *Le Développement des Quantités Chez l'Enfant*. Neuchatel: Delachaux et Niestlé.

Piaget, J., Inhelder, B., & Szeminska, A. (1948). *La Géométric Spontanée de l'Enfant*. Paris: Presses Universitaires de France.

Piaget, J., & Szeminska, A. (1941). *La Genèse du Nombre Chez l'Enfant*. Neuchatel: Delachaux et Niestlé.

Quinton, G., & Fellows, B. (1975). "Perceptual" strategies in the solving of three-term series problems. *British Journal of Psychology, 66*, 69–78.

Reckase, M.A. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 271–286). New York: Springer-Verlag.

Reyna, V.F., & Brainerd, C.J. (1990). Fuzzy processing in transitivity development. *Annals of Operations Research, 23*, 37–63.

Roussos, L.A., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1–30.

Scheiblechner, H. (1972). Das lernen un lösen complexer Denkaufgaben (Learning and solving complex thought problems). *Zeitschrift für experimentelle und angewandte Psychologie, 19*, 481–520.

Sijtsma, K., & Meijer, R.R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149–157.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage Publications.

Smedslund, J. (1963). Development of concrete transitivity of length in children. *Child Development, 34*, 389–405.

Smedslund, J. (1969). Psychological diagnostics. *Psychological Bulletin, 71*, 237–248.

Sternberg, R.J. (1980a). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology, 109*, 119–159.

Sternberg, R.J. (1980b). The development of linear syllogistic reasoning. *Journal of Experimental Child Psychology, 29*, 340–356.

Sternberg, R.J., & Weil, E.M. (1980). An aptitude $\times$ strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology, 72*, 226–239.

Stout, W. (1993). *DIMTEST*. Urbana-Champaign, IL: The William Stout Institute for Measurement.

Stout, W. (1996). *DETECT*. Urbana-Champaign, IL: The William Stout Institute for Measurement.

Stout, W., Froelich, A.G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A.Boomsma, M.A.J. van Duijn, & T.A.B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 357–375). New York: Springer.

Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L.A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331–354.

Thomas, H., Lohaus, A., & Kessler, T. (1999). Stability and change in longitudinal water-level task performance. *Developmental Psychology, 35*, 1024–1037.

Trabasso, T. (1977). The role of memory as a system in making transitive inferences. In R.V. Kail, J.W. Hagen, & J.M. Belmont (Eds.), *Perspectives on the Development of Memory and Cognition* (pp. 333–366). Hillsdale, NJ: Erlbaum.

Trabasso, T., Riley, C.A., & Wilson, E.G. (1975). The representation of linear order and spatial strategies in reasoning: a developmental study. In R.J. Falmagne (Ed.), *Reasoning: Representation and Process in Children and Adults* (pp. 201–229). Hillsdale, NJ: Erlbaum.

Van Abswoude, A. A.H., Van der Ark, L.A., & Sijtsma, K. (2004). A comparative study on test dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3–24.

Verweij, A.C. (1999). *Scaling transitive inference in 7–12 years old children*. Unpublished doctoral dissertation, Vrÿe Universiteit Amsterdam, the Netherlands.

Verweij, A.C., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development, 23*, 241–264.

Wright, B.C. (2001). Reconceptualizing the transitive inference ability: A framework for existing and future research. *Developmental Review, 21*, 375–422.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129–152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–249.

Manuscript received 12 FEB 2003
Final version received 17 FEB 2004